MthStat 465, Spring 2005, Lecture Number 4// Median-Median Line and Least Squares Lines

The use of Excel was discussed, and illustrated by two methods of fitting a line to data. In what follows, the data is assumed to be bivariate, that is, ordered pairs $((x_1, y_1), \dots, (x_N, y_N))$.

1. Median-Median Line

The data is broken up into 3 roughly equal groups by arranging the ordered pairs from smallest to largest in lexicographical order. This means consider the first coordinates first when deciding order. Then for each group assign a representative point whose first coordinate is the median of the first coordinates of the group and whose second coordinate is the median of the second coordinates of the group. Use the representatives of the high and low groups to determine the slope of the line, m. Then, using the point slope formula, write the point slope form of a line through each point, using slope m. Then average the equations. In terms of formulae: Let (x_L, y_L), (x_M, y_M), and (x_H, y_H) denote the representative points for the low, middle and high groups respectively. Let $m = (y_H - y_L)/(x_H - x_L)$. We have three equations (which two are actually the same?):

$$\begin{aligned} y - y_L &= m(x - x_L) \\ y - y_M &= m(x - x_M) \\ y - y_H &= m(x - x_H) \end{aligned}$$

The equation of the Median-Median line is

$$\frac{1}{3}\left((y - y_L) + (y - y_M) + (y - y_H)\right)\frac{1}{3}\left(m(x - x_L) + m(x - x_M) + m(x - x_H)\right).$$

If we put

$$\begin{aligned} x_0 &= \frac{x_L + x_M + x_H}{3} \\ y_0 &= \frac{y_L + y_M + y_H}{3} \end{aligned}$$

we see that the equation of the Median-Median line can be written

$$y - y_0 = \frac{y_H - y_L}{x_H - X_L}(x - x_0).$$

The point (x_0, y_0) is the centroid of the triangle with vertices at (x_L, y_L), (x_M, y_M), and (x_H, y_H). For details and numerical examples, see

`www.math.uncc.edu/~droyster/courses/spring00/maed3103/Median-Median_Line.htm`

2. Least Squares

In the method of least squares, a numerical measure of how well a curve fits the data is given. A curve is then chosen to minimize this measure. Specifically, if $y = f(x)$ is proposed as a fit to the data, then

$$E(f) := \sum_{k=1}^{N} (f(x_k) - y_k)^2$$

measures how well $y = f(x)$ fits the data: the bigger $E(f)$ is, the worse the fit. $E(f)$ is called the sum of squared errors. The goal of the method of least squares is to choose f so that $E(f)$, the **sum of squared errors**, is as small as possible. This minimum value is called the **residual sum of squares**. Usually there is a restriction on the choices for f. The most common is that the graph of $y = f(x)$ must be a line. In the case of lines we let $\overline{x}$ and $\overline{y}$ be the averages of the first and second coordinates of the data respectively. We then consider all functions f of the form

$$f(x) = m(x - \overline{x}) + y + e.$$

If we let m and E vary over all real numbers, $y = f(x)$ will describe all possible non-vertical lines in the coordinate plane. $E(f)$ is then a non-negative quadratic function of m and e. It is minimized by taking $e = 0$ and

$$m = \frac{\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{N}(x_k - \overline{x})^2}.$$

The residual sum of squared errors is

$$\sum_{k=1}^{N}(y_k - \overline{y})^2 - \frac{m^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2}.$$

Note then, that the least squares line of best fit has the form

$$y - \overline{y} = \frac{\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{N}(x_k - \overline{x})^2}(x - \overline{x}).$$

For details, see *A painless approach to least squares*, handed out in class.